

**Sample Size for a Phylogenetic Inference**

by

**Gary A. Churchill,**

**Arndt von Haeseler**

and

**William C. Navidi**

**BU-1125-MA**

**October 1991**

# Sample Size for a Phylogenetic Inference

Gary A. Churchill<sup>1</sup>, Arndt von Haeseler<sup>2</sup> and William C. Navidi<sup>2</sup>

<sup>1</sup>Department of Plant Breeding and Biometry  
Cornell University

<sup>2</sup>Department of Mathematics  
University of Southern California

Address for Correspondence:

Biometrics Unit, 337 Warren Hall  
Cornell University  
Ithaca, NY 14853

Running head: Sample Size for a Phylogenetic Inference

Abbreviations:

MLE: maximum likelihood estimate

LRT: likelihood ratio test

df: degrees of freedom

## Abstract

The objective of this work is to describe sample size calculations for the inference of a non-zero central branch length in an unrooted four species phylogeny. Attention is restricted to independent binary characters, such as might be obtained from an alignment of the purine-pyrimidine sequences of a nucleic acid molecule. A statistical test based on a multinomial model for character state configurations is described. The importance of including invariable sites in models for sequence change is demonstrated and their effect on sample size is quantified. The methods are applied to a four species alignment of small subunit rRNA sequences derived from two archaeobacteria, a eubacteria and a eukaryote. We conclude that the information in these sequences is not sufficient to resolve the branching order of this tree. Estimates of the number of aligned nucleotide positions required to provide a reasonably powerful test are given.

Keywords: phylogenetic inference, invariable sites, multinomial model

## Introduction

In Felsenstein (1988), methods of phylogenetic inference are reviewed with an emphasis on reliability of the procedures. In the past, much emphasis has been placed on finding methods to compute the best tree with little emphasis on the confidence with which we can say that the tree is correct. One of the standard arguments given to support an inferred tree is that the sample size, the number of aligned nucleotides in the sequence set, is sufficiently large. This is based on the common sense and statistically well founded idea that increasing the sample size should increase the reliability of the reconstruction. However, it is well known that tree reconstruction by the method of maximum parsimony can fail to give the correct tree even when the sample size is large (Felsenstein 1978). Maximum likelihood methods are consistent. They are guaranteed to give the correct reconstruction provided the sample size is large and the model assumptions are true. Nevertheless, it remains an open question how large the sample size has to be before one can conclude that a given tree is correct.

In this paper, we will focus on inferences concerning the interior branch of a four species unrooted tree. The data consist of 4 binary  $(0, 1)$  sequences with known alignment. We use a simple model with symmetric mutation rates which is essentially that of Cavender (1978), except that we allow a proportion of the sites to be invariable. Under the assumptions of

independence and identical substitution rates across sites, we are led to consider a multinomial model for binary character state configurations. The model is described in detail below. A model-based test for the star phylogeny versus a tree phylogeny is developed. Examination of the asymptotic power properties of this test will lead us to consideration of the sample size required to make specific inferences about the structure of a phylogenetic tree. The methods are applied to a problem of current interest concerning the origin of archaeobacteria (e.g. Woese 1987, Lake 1988).

## The Model

**Topologies** The evolutionary relationships among a set of sequences can be described by a graph structure. The terminal nodes of the graph represent the observed sequences and interior nodes represent hypothetical ancestral sequences.

If no pair of sequences share more common history than any other pair, the relationships will be described best by a star phylogeny (figure 1a). In the rooted case, where the ancestral sequence can be placed unambiguously on one of the branches, this corresponds to a trifurcation. Here we will consider the unrooted case. Without loss of generality we place the root at the central node and assume that all four sequences have evolved independently from a common ancestor whose sequence is the root.

If two of the sequences have shared a common evolutionary history, the relationship is best described by a phylogeny which is a bifurcating tree. For much of our discussion we will focus on the particular tree (topology 1) that splits the pair 1,2 from the pair 3,4 (figure 1b). The root is placed at the node labeled (\*) and the branches are numbered as shown. Two other tree topologies are possible. We will refer to the tree that splits the pairs 1,3 from 2,4 as topology 2 and the tree that splits 1,4 from 2,3 as topology 3.

**The Substitution Process** We suppose that at some initial time all 4 sequences had a common ancestor and that they have since diverged by a substitution process to their present state. We will make the assumptions that the substitution process is independent and identical at each site and further that substitutions are equally likely from 1 to 0 and from 0 to 1. In the case of nucleic acid sequences, 0 and 1 will usually represent purine and pyrimidine. Substitution events that exchange purines and pyrimidines are called transversions and they generally occur at a lower rate than transitions, changes that preserve purine-pyrimidine states (Nei 1987, p. 28).

The parameters of our model are the probabilities that a particular site differs in the two sequences represented by the endpoints of a branch. Let  $\theta_i$  denote the probability that at a given site an odd number of substitutions have occurred along branch  $i$  and thus that the endpoints differ. If substitution events occur as a Markov process along each branch then it follows that the parameter  $\theta_i$  is bounded above by 1/2. The value of  $\theta_i$  may be considered a measure of evolutionary distance and we will refer to it as the length of branch  $i$ .

**Invariable Sites** Many nucleotide sequences are involved in essential functions of organisms. Hence, it is likely that for some sites substitution events will not happen. The concept of invariable sites was introduced by Fitch and Margoliash (1967). In general it will not be possible to distinguish invariable sites from variable sites which by chance are

unvaried. We will assume that any invariable sites are invariable in all four sequences, although this may not be true in general (Fitch 1971). We can extend our model by introducing one additional parameter,  $\theta_0$ , which is the probability that a given site is invariable. Notice that the introduction of invariable sites does not violate the identically distributed sites assumption. Any randomly selected site will be invariable with probability  $\theta_0$  and otherwise variable.

**Character State Configurations** At each site in the set of aligned sequences we observe an ordered quartet of zeroes and ones, a binary character state configuration. The basic data for our analysis will be a vector of configuration counts,  $\mathbf{x}$ , with components

$$\begin{aligned}x_0 &= \langle 0000 \rangle + \langle 1111 \rangle \\x_1 &= \langle 1000 \rangle + \langle 0111 \rangle \\x_2 &= \langle 0100 \rangle + \langle 1011 \rangle \\x_3 &= \langle 0010 \rangle + \langle 1101 \rangle \\x_4 &= \langle 0001 \rangle + \langle 1110 \rangle \\x_5 &= \langle 1100 \rangle + \langle 0011 \rangle \\x_6 &= \langle 1010 \rangle + \langle 0101 \rangle \\x_7 &= \langle 1001 \rangle + \langle 0110 \rangle\end{aligned}\tag{1}$$

The bracket notation  $\langle abcd \rangle$  represents the total number across all sites of occurrences of the character state configuration  $abcd$ .



When the sites are independent and identically distributed, the distribution of  $\mathbf{x}$  will be multinomial with probability vector  $\mathbf{p}$ . Without any structural assumptions, the multinomial parameter  $\mathbf{p}$  can take any value in the simplex  $S^8 = \{\mathbf{p} : p_i > 0, \sum_{i=1}^8 p_i = 1\}$ , the set of all possible values for a probability vector with 8 components. When the structure of an evolutionary tree model is assumed, the values that  $\mathbf{p}$  can take are constrained to lie in a subset of  $S^8$ . The dimension of this subset is defined to be the number of free parameters in the model. Its form is determined by the topology of the tree which relates the sequences.

**Configuration Probabilities** In this section we will define the configuration probabilities as functions of the parameter vector  $\boldsymbol{\theta} = (\theta_0, \dots, \theta_5)$ , i.e.,  $\mathbf{p} = \mathbf{p}(\boldsymbol{\theta})$ . We begin with the definition of  $\mathbf{p}(\boldsymbol{\theta})$  for a star phylogeny, generalize to tree phylogenies and then allow for invariable sites.

**Star Phylogeny** Under the star phylogeny model of sequence evolution, the configuration probabilities are functions of  $\theta_1, \dots, \theta_4$  and will be denoted by  $p_i(\boldsymbol{\theta}) = b_i$ , where

$$b_0 = (1-\theta_1)(1-\theta_2)(1-\theta_3)(1-\theta_4) + \theta_1\theta_2\theta_3\theta_4$$

$$b_1 = \theta_1(1-\theta_2)(1-\theta_3)(1-\theta_4) + (1-\theta_1)\theta_2\theta_3\theta_4$$

$$b_2 = (1-\theta_1)\theta_2(1-\theta_3)(1-\theta_4) + \theta_1(1-\theta_2)\theta_3\theta_4$$

$$b_3 = (1-\theta_1)(1-\theta_2)\theta_3(1-\theta_4) + \theta_1\theta_2(1-\theta_3)\theta_4$$

$$b_4 = (1-\theta_1)(1-\theta_2)(1-\theta_3)\theta_4 + \theta_1\theta_2\theta_3(1-\theta_4)$$

$$\begin{aligned}
b_5 &= \theta_1 \theta_2 (1-\theta_3)(1-\theta_4) + (1-\theta_1)(1-\theta_2) \theta_3 \theta_4 \\
b_6 &= \theta_1 (1-\theta_2) \theta_3 (1-\theta_4) + (1-\theta_1) \theta_2 (1-\theta_3) \theta_4 \\
b_7 &= \theta_1 (1-\theta_2)(1-\theta_3) \theta_4 + (1-\theta_1) \theta_2 \theta_3 (1-\theta_4)
\end{aligned} \tag{2}$$

If the ancestral state at site  $i$  is zero, the left and right terms to the right of the equals sign in equation (2) are the probabilities of the left and right terms to the right of the equals sign in equation (1). If the ancestral state at site  $i$  is one, this correspondence is reversed.

**Tree Phylogenies** Under a tree phylogeny model, the central branch parameter  $\theta_5$  is no longer constrained to be zero. The meaning of  $\theta_5$  and the functional form of the configuration probabilities depends on the topology of the tree. We will define the following permutations on the indices  $\{0, \dots, 7\}$ , of the vectors  $\mathbf{p}$  and  $\mathbf{x}$ :

$$\begin{aligned}
\pi_1 &= \{5, 2, 1, 4, 3, 0, 7, 6\} \\
\pi_2 &= \{6, 3, 4, 1, 2, 7, 0, 5\} \\
\pi_3 &= \{7, 4, 3, 2, 1, 6, 5, 0\}
\end{aligned} \tag{3}$$

These permutations show, for each of the tree topologies (1, 2 and 3), the correspondence between configurations that differ by an odd number of substitutions on the central branch. For example, assuming topology 1, a substitution along the central branch will change configuration 0 to configuration 5. Thus  $\pi_1(0) = 5$ . Similarly, a central branch substitution will change configuration 3 to configuration 4. Thus  $\pi_1(3) = 4$ . The

permutations allow us to express the configuration probabilities for any tree topology  $k$  ( $k = 1, 2, 3$ ) as functions of the star phylogeny probabilities  $b_i$  and the central branch parameter  $\theta_5$ . The configuration probabilities for tree topology  $k$  are  $p_i(\theta) = f_i$ , where

$$f_i = (1 - \theta_5)b_i + \theta_5 b_{\pi_k(i)} . \quad (4)$$

Thus, the different tree topologies generate configuration probabilities that are distinct linear combinations of the star phylogeny probabilities weighted by the central branch parameter  $\theta_5$ .

**Invariable Sites** We introduce invariable sites by allowing any given site along the aligned sequences to be invariable with probability  $\theta_0$ . An invariable site is not subject to substitution in any branch of the tree. Configuration probabilities are given by  $p_i(\theta) = g_i$ , where

$$\begin{aligned} g_0 &= \theta_0 + (1 - \theta_0)f_0 \\ g_i &= (1 - \theta_0)f_i; \quad i = 1, \dots, 7. \end{aligned} \quad (5)$$

Thus, the effect of including invariable sites is to inflate the probability of the character state configuration  $x_0$ . A star phylogeny model with invariable sites can be obtained by restricting the value of  $\theta_5$  to be zero.

## Methods

In this section, we give a brief description of the statistical methods that are relevant to our phylogenetic inference problem. A general treatment of inference for parametric multinomial models can be found in Rao (1973, pp. 359-363). The likelihood approach is essentially the same as that of Felsenstein (1981).

A maximum likelihood estimate (MLE) of the model parameter is obtained by maximizing the log-likelihood function

$$\ell(\theta) = \sum_{k=0}^7 x_k \ln(p_k(\theta)) \quad (6)$$

over the range of allowable values for  $\theta$ . Values consistent with our model are  $(0, \frac{1}{2})$  for  $\theta_1, \dots, \theta_5$  and  $(0, 1)$  for  $\theta_0$ . Maximization over a wider range is allowed, provided that the estimated probabilities  $p(\hat{\theta})$  lie within the simplex  $S^8$ . In this case estimates of  $\theta_i$  ( $i = 1, \dots, 5$ ) may fall outside the interval  $(0, 1/2)$ . Small negative estimates can usually be attributed to random variation around a true value near zero. However, significantly large negative estimates suggest some failure of the model assumptions. Estimates greater than  $1/2$  are unlikely to occur in practice.

We wish to test the null hypothesis of a star phylogeny,  $H_0: \theta_5 = 0$  against one of three possible alternative hypotheses. (In this notation for hypotheses, the colon may be read as "that".) The alternative hypothesis assumes a particular tree phylogeny with non-zero central branch length  $H_A: \theta_5 \neq 0$ . Notice that we have chosen to consider a two-sided

alternative. Although we do not expect negative values for  $\theta_5$  to be “true”, allowing negative estimates helps us to avoid problems that can occur for inferences on the boundary of a parameter space. The result is a slightly conservative test, because we allow a wider range of alternatives. The advantage is that standard results for the large-sample distribution of test statistics apply.

Let  $\hat{\theta}$  denote the MLE of the vector  $\theta$  under the alternative hypothesis and let  $\tilde{\theta}$  denote the MLE of  $\theta$  under the null hypothesis. The likelihood ratio statistic for testing  $H_0$  versus  $H_A$  is

$$L = 2(\ell(\hat{\theta}) - \ell(\tilde{\theta})) \quad . \quad (7)$$

Under the null hypothesis,  $L$  has a large sample distribution which is approximately chi-squared with 1 degree of freedom (df). When the alternative hypothesis is true, the large sample distribution of  $L$  is non-central chi-squared with 1 df and non-centrality parameter  $\delta$ . The quantity  $\delta$  depends both on the sample size and the specific value of  $\theta$  which is assumed to be true. See Appendix A for details on the computation of  $\delta$ .

The generalized likelihood ratio statistic  $G^2$  is defined to be twice the difference between the maximum value of the log-likelihood attained over  $S^8$  and the maximum log-likelihood attained under the model constraints. This statistic is described by Navidi *et al.* (1991) in the context of phylogenetic inference and in a more general setting as a measure of goodness-of-fit by Bishop *et al.* (1975, pp. 125-130). When the specified model is correct,  $G^2$  will have an approximate chi-squared distribution

with degrees of freedom equal to seven minus the number of free parameters in the model. The statistic  $L$  is used to compare two different models e.g., star phylogeny versus a tree phylogeny. The statistics  $G^2$  is used to compare a model to the maximum log-likelihood without model constraints. Likelihood ratio statistics for comparing two appropriately nested models can be computed by taking the difference in their respective  $G^2$  values. Log-likelihoods have also been used to compare non-nested models (e.g. Hasegawa and Kishino, 1989).

A test with type I error probability  $\alpha$  is given by the following procedure:

$$\text{Reject } H_0 \text{ when } L > \chi^2_{\alpha,1}, \quad (8)$$

where  $\chi^2_{\alpha,1}$  denotes the upper  $\alpha$  percentile of a chi-squared distribution on 1 df. We denote the type II error probability of a test by  $\beta$  and define the power to be  $1 - \beta$ . When a specific value of  $\theta$  under the alternative hypothesis is true, the power of the test is  $\Pr(L > \chi^2_{\alpha,1})$  where  $L$  has a non-central chi-squared distribution as described above.

A typical sample size calculation might proceed as follows. We have in mind a specific alternative,  $H_A : \theta = (\theta_0, \dots, \theta_5)$ . If the alternative is true, we wish to reject the null hypothesis with probability at least  $1 - \beta$  using a level  $\alpha$  test. A sample size is chosen and the power of the test is computed. If the power is too small, the sample size should be adjusted upward. If the power is greater than  $1 - \beta$  a smaller sample size can be considered.

There are often considerations other than the power of a test that limit the sample size. More data may be too expensive to collect or may simply not be available. In this situation, the sample size is fixed and the power can be computed against a number of alternatives. This will give the investigator a feeling for the range of alternatives that can be supported by the available data.

## Results

**Power Curves** Power, the probability that you will not fail to reject a false null hypothesis, for the test of  $H_0 : \theta_5 = 0$  was computed for several arrangements of branch lengths and a range of sample sizes. (See Appendix A for details of calculations.) The results are summarized in Figures 2 and 3 in order to illustrate the following points.

- (1) For fixed values of  $\theta_1, \dots, \theta_4$  and  $n$ , power increases with increasing  $\theta_5$ .
- (2) For fixed values of  $\theta_1, \dots, \theta_5$ , power increases with increasing  $n$ .
- (3) Power of the test decreases as the proportion of invariable sites increases.
- (4) For fixed values of  $n$  and  $\theta_5$ , power decreases as any of the outer branch lengths are increased.
- (5) For trees with asymmetric outer branches, the better power is achieved when the two shortest branches are separated by the central branch.

Figure 2a shows the power of the test as a function of  $\theta_5$  for a tree with symmetric outer branch lengths and no invariable sites. Figure 2b shows power as a function of  $\theta_0$  for a fixed central branch length. The results are approximate and, for large values of  $\theta_0$ , the true power is somewhat less than shown. The loss of power with increasing  $\theta_0$  is such that if a sample of size  $n$  gives a test with power  $1 - \beta$  when  $\theta_0 = 0.0$ , a



sample of size  $n/(1-\theta_0)$  will produce a test with the same power, all other parameters being constant. This is a consequence of the form of the Fisher's information matrix under a model with invariable sites.

Figure 3 shows the power of the test as a function of sample size for a range of central branch lengths. These curves give better resolution for parameter values in the range of interest, i.e. small central branch lengths. Figures 3 a, b and c show the power curves for trees with symmetric outer branches of increasing length. The sample sizes needed to achieve the same power are seen to increase by as much as 2 orders of magnitude for these parameter values. The same qualitative results are obtained when any subset of outer branch lengths is increased.

Figures 3d and e show power curves for trees with asymmetric outer branch lengths. In 3d the shorter branches are adjacent, and in 3e the shorter branches are separated by the central branch. The better power is achieved when the two shorter branches of the tree are separated by the central branch. This observation is intuitively clear when we consider that most of the information about central branch events comes from the sequences with the fewest changes in the outer branches. The values 0.1 and 0.3 for outer branch length parameters were chosen to be consistent with the example below. When the central branch length is  $\theta_5 = 0.1$ , these trees correspond to the borderline case where parsimony methods are more likely to select the wrong tree (Felsenstein, 1978).

**Simulations** The results stated above are based on asymptotic approximations which are valid for large samples. To check the quality of these approximations, limited simulations were conducted. The results are plotted as symbols superimposed on the curves in figure 2a. Each symbol represents 1000 simulations. The level  $\alpha$  of the likelihood ratio test (i.e. power at  $\theta_5=0.0$ ) is very close to  $\alpha = 0.05$ , even for the smallest sample sizes considered. The power of the tests agrees very well with asymptotic theory for sample sizes greater than 1000bp, over a wide range of parameter values. The asymptotic theory is also very accurate for short central branch lengths. The actual power is somewhat less than predicted when the sample size is small and the central branch length is large.

**Comparison With Other Methods** Nucleic acid sequences have 4 distinct states which we have collapsed to a binary purine-pyrimidine classification. A general model for 4-state sequences is described by Navidi *et al.* (1991). The dimensions of  $\mathbf{x}$  and  $\mathbf{p}$  are increased to  $4^4 = 256$  and the number of free parameters needed to specify the model is 63 (plus one if we allow for invariable sites). The central branch test becomes a test for a  $4 \times 4$  transition matrix being equal to the identity matrix. Hence, the likelihood ratio statistic is approximately chi-squared with 12 df, the number of free parameters in a  $4 \times 4$  transition matrix. Power calculations are a straightforward generalization of the binary case (see Appendix A).

For purposes of illustration, we have chosen a Jukes-Cantor like model for 4-state sequence evolution. The off-diagonal elements of the matrix for branch  $i$  were set equal to  $\theta_i/2$  to give probabilities of a transversion that are comparable to the binary models. Power for testing the central branch hypothesis was computed for the asymmetric tree models with the shorter branches together or apart and a sample size of  $n=1000$ . The full 63 or 64 parameter model was assumed to hold in the 4-state case. The curves shown in Figure 4 for binary and 4-state models with and without invariable sites are roughly comparable. Thus, in this case the power of the tests using binary data or the full 4-state data are similar. Because a wide range of 4-state models could be considered, we make no general conclusions.

Linear invariant statistics provide methods for testing hypotheses about evolutionary trees that can be robust to many of the assumptions required for likelihood analysis. In particular these methods are valid when rates of sequence change vary across sites, i.e. sites are not identically distributed. What is the cost of using linear invariant methods when the likelihood ratio test assumptions are in fact valid? To answer this question, we have computed the power of a test based on Lake's invariants (Lake, 1987). See Appendix B for details. The loss of power relative to the LRT is dramatic. With a sample size of 1000, we have only 30% power against an alternative  $\theta_5 = 0.1$ .

### An Example: The Archaeobacteria Tree

To provide an example of the methods discussed above, we extracted four aligned sequences from the collection of small subunit rRNAs (Dams *et al.* 1988). The sequences are

1. *Sulfolobus solfataricus*, an archaeobacteria,
2. *Halobacterium salinarium*, an archaeobacteria,
3. *Escherichia coli*, a eubacteria,
4. *Homo sapiens*, a eukaryote.

There are a total of 1352 aligned positions. The question of interest with respect to these sequences is whether the archaeobacteria are monophyletic or polyphyletic in origin. The monophyletic archaeobacterial tree is tree 1 in our notation. We wish to determine which, if any, of the three trees is supported by the data. The configuration count vector is  $\mathbf{x} = (787, 45, 56, 145, 178, 68, 33, 40)$ .

Table 1 summarizes the models that were fit to these data. Likelihood ratio test statistics can be computed by taking differences in  $G^2$  values. For example, to test for a non-zero central branch length under the model of tree 1 with no invariable sites, we compute the difference  $L = 23.60 - 15.47 = 8.13$ . The probability that a chi-squared random variable with 1 degree of freedom will exceed this value is  $p < 0.005$ . This is strong evidence in favor of a non-zero central branch, provided that the model assumptions are correct. In fact, when invariable sites are not included in

the model, evidence in favor of each of the central branches is strong. However, in all cases, the goodness-of-fit is poor as indicated by the significantly large values of  $G^2$ . This observation confirms the claim of Shoemaker and Fitch (1989) that invariable sites are necessary (although perhaps not sufficient) to obtain a reasonable fit to most data on nucleic acid sequence divergence.

The simplest model that provides an adequate fit to the data is the star phylogeny with invariable sites ( $G^2=0.96$  on 2df,  $p=0.62$ ). The evidence in favor of invariable sites is strong. Comparing the star phylogenies with and without invariable sites, we have  $L = 22.64$  ( $p<.0001$ ). None of the three tree phylogenies can significantly improve upon the fit obtained with the star phylogeny.

Parameter estimates obtained under the various models are summarized in table 1. Notice that the estimated outer branch lengths,  $\hat{\theta}_1, \dots, \hat{\theta}_4$  increase sharply when invariable sites are added to the model and decrease only slightly when a central branch is added. Standard errors for the parameter estimates  $\hat{\theta}_0, \dots, \hat{\theta}_4$  under the star model with invariable sites are 0.0468, 0.0155, 0.0171, 0.0258 and 0.0285 respectively. The estimated proportion of invariable sites is 30%. However, the uncertainty in this estimate is large, with 95% confidence interval (0.201, 0.3844).

The power of the tests for non-zero central branch lengths under any of the three tree topologies is low. Assuming outer branch lengths  $\theta_1 = \theta_2$

$= 0.1$  and  $\theta_3 = \theta_4 = 0.3$  and a proportion of invariable sites  $\theta_0 = 0.3$ , the power for testing against the alternative tree 1 with central branch  $\theta_5 = 0.02$  is only  $1 - \beta = 0.066$ . The power for testing against the alternative tree with the shorter branches separated by the central branch and  $\theta_5 = 0.02$  is only slightly better at  $1 - \beta = 0.121$ . We conclude that there is insufficient information in these data to resolve the branching order of the four species.

Sample sizes required to achieve 60% and 80% power against the null hypothesis are summarized in table 2 for a variety of central branch lengths and two distinct tree topologies, assuming there are no invariable sites. Sample sizes for models with invariable sites can be obtained by rescaling the sample size values  $n$  to be  $n/(1-\theta_0)$ . The least favorable situation for inferences concerning the central branch length is topology 1. For a central branch length  $\theta_5 = 0.02$  and 30% invariable sites, a sample size of 78,000 is needed to obtain a reasonably powerful test ( $1-\beta=80\%$ ). The situation is somewhat better for topologies 2 and 3, where a sample size of 18,000 would provide 80% power against the same alternative.

## Discussion

Phylogenetic inference in the context of a multinomial model for character state configurations has the advantage of bringing a large body of powerful inference procedures to bear on the problem. However, a major drawback is that the model assumptions can impose unrealistic constraints. The importance of including an invariable sites parameter has been demonstrated. This is a first step towards more realistic models which would allow rates of sequence change to vary across sites. Including additional parameters in a model when they are not needed decreases our power to make inferences. However, failure to include relevant parameters results in a misspecified model and can produce misleading results. Methods exist that are robust to constant rate assumptions, namely linear invariant methods, but the price for robustness is a loss of power. There is clearly a tradeoff between the power to make inferences and the generality of the model assumptions. The discrepancy in power between multinomial tests and invariant based tests suggests that a wide range of intermediate results could be obtained. The explicit modeling of rate variation would seem to be a worthwhile pursuit.

In our example, using binary data seems to capture much of the relevant information. However, for less divergent sequences, it may be necessary to use the full 4-state data. Large sample sizes may be required to estimate all of the parameters accurately. The number of parameters

can be effectively reduced by modelling. For example, Felsenstein (1981) proposed a model which requires only 17 free parameters. The importance of such models increases when we consider trees with more than 4 species. However, their realism should be carefully considered.

The question remains, how can we improve the power of the central branch inference in our example given the limited size of rRNA sequences? As we observed, in the results above, bringing the outer nodes closer to the central branch can increase the power dramatically. This suggests that the best species to consider are those with the least divergence, i.e. the slowest rates of evolution. See Lake (1990) for an application which illustrates this point. An alternative means of bringing the "outer" nodes closer is to include more species in the tree. The maximum benefit would be obtained by including species with the earliest divergence times. There is, however, a tradeoff in that we have more parameters to estimate. Adding a few carefully selected species may prove to be the best approach.

The power calculations for a 4 species tree are still relevant. If the "outer" nodes are actually root nodes of an extended phylogeny, our knowledge of the sequences represented by these nodes is not precise. A sequence can be inferred and is represented as a probability at each nucleotide position of being 0 or 1. This inferred sequence is readily available when the tree likelihood has been maximized using an EM algorithm (Felsenstein 1988). The configuration "counts" are still sufficient statistics for the central branch inference but may not be



integers. Power calculations assuming perfect knowledge of these node sequences will lead to a lower bound for the sample size.

## Appendix A: Computing the non-centrality parameter

Fisher's information matrix for a parametric multinomial model with  $s$  outcome states has elements

$$I_{ij}(\theta) = \sum_{k=1}^s \frac{1}{p_k(\theta)} \frac{\partial}{\partial \theta_i} p_k(\theta) \frac{\partial}{\partial \theta_j} p_k(\theta) \quad (9)$$

where the indices  $i$  and  $j$  run through the dimensions of  $\theta$ . The parameter vector can be partitioned as

$$\theta = (\theta_1, \theta_2)' \quad (10)$$

so that the null hypothesis becomes  $H_0 : \theta_1 = 0, \theta_2 = \text{anything}$ . The dimension of  $\theta_1$  is  $r$  and the dimension of  $\theta_2$  is  $s-r$ . The information matrix is partitioned to correspond with the partitioning of  $\theta$  as

$$I(\theta) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} . \quad (11)$$

We define

$$I_{11.2} = I_{11} - I_{12} I_{22}^{-1} I_{21} , \quad (12)$$

to be the information for  $\theta_1$  corrected for having estimated  $\theta_2$ .

A specific value of  $\theta$  under the alternative hypothesis is assumed to be true. The non-centrality parameter is then given by

$$\delta^2 = \frac{n}{2} \theta_1' I_{11.2} \theta_1 . \quad (13)$$

Tables of the non-central chi-squared distribution, or a simple computer algorithm (Posten, 1989) can be used to find the probability that a non-central chi-squared random variable on  $r$  df with non-centrality  $\delta$  will exceed the upper  $\alpha$  quantile of a central chi-squared distribution.

## Appendix B: Power Calculations For an Invariants Test.

A one sided, exact version of Lake's procedure (Lake, 1987) for selecting among alternative tree topologies has been proposed by Navidi *et al.* (1991). We have adopted the following notation from Navidi *et al.* (1991): Let  $\mathbf{v}_i (i=1,2,3)$  be the 256 dimensional vector with ones in positions corresponding to the positive terms of Lake's invariant statistic for tree  $i$  and zeros elsewhere and let  $\mathbf{w}_i (i=1,2,3)$  be the vector with ones in positions corresponding to the negative terms and zeros elsewhere. Let  $\mathbf{x}$  be the 256 dimensional vector of configuration counts and let  $\mathbf{p}$  be the corresponding vector of probabilities. Define  $P_i = \mathbf{v}_i' \mathbf{x}$  and  $Q_i = \mathbf{w}_i' \mathbf{x}$ . (Prime denotes the transpose of a column vector to a row vector.) Under the independent and identically distributed assumptions,  $P_i + Q_i$  is binomial on  $n$  with probability  $(\mathbf{v}_i + \mathbf{w}_i)' \mathbf{p}$  and the conditional distribution of  $P_i$  given  $P_i + Q_i = k$  is binomial on  $k$  with probability

$$\pi = \mathbf{v}_i' \mathbf{p} / (\mathbf{v}_i + \mathbf{w}_i)' \mathbf{p} . \quad (14)$$

When tree  $i$  is correct,  $\pi = \frac{1}{2}$ .

The procedure is to reject the null hypothesis,  $H_0$  : tree  $i$  is not correct, in favor of the alternative,  $H_A$  : tree  $i$  is correct, if  $P_i > p_{1-\alpha}$ , where  $p_{1-\alpha}$  is the  $1 - \alpha$  quantile of the binomial distribution on  $P_i + Q_i$  with probability  $\frac{1}{2}$ . The power of this test is

$$\Pr(\text{Reject}) = \sum_{k=0}^n \Pr(\text{Reject} | P_i + Q_i = k) \Pr(P_i + Q_i = k) \quad (15)$$

where

$$\Pr(\text{Reject} | P_i + Q_i = k) = \sum_{j=p_{1-\alpha}}^k \binom{k}{j} \pi^j (1-\pi)^{k-j} . \quad (16)$$

### **Acknowledgments**

This research was supported in part by the National Science Foundation (grant DMS90-05833). We would like to express our appreciation to Simon Tavaré and to Walter Fitch for their valuable discussions and careful readings of the manuscript.

Literature Cited

- Bishop, Y.M.M., S.E. Feinberg and P.W. Holland 1975. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge MA.
- Cavender, J.A. 1978. Taxonomy with confidence. *Math. Biosciences* 40:271-280 (Erratum 44:309, 1979.).
- Dams, E., and L. Hendriks, Y. Van de Peer, et al. 1988. Compilation of small subunit RNA subsequences. A supplement to *Nucleic Acids Research* 16:87-174.
- Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27:401-410.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17:368-376.
- Felsenstein, J. 1988. Phylogenies from Molecular Sequences: Inference and Reliability. *Annual Review of Genetics* 22:521-565.
- Fitch, W.M. 1971. The Nonidentity of Invariable Positions in the Cytochromes c of Different Species. *Biochem. Gen.* 5:231-241.
- Fitch, W.M., and E. Margoliash. 1967. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. *Biochem. Gen.* 1:65-71.
- Hasegawa, M. and H. Kishino. 1989. Confidence limits on the maximum likelihood estimate of the hominoid tree from mitochondrial DNA sequences. *Evolution.* 43:672-677.

- Lake, J.A. 1987. A rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Mol. Biol. and Evol.* 4:167-191.
- Lake, J.A. 1988. Origin of the eukaryotic nucleus determined by rate invariant analysis of rRNA sequences. *Nature* 331:184-186.
- Lake, J.A. 1990. Origin of the Metazoa. *Proc. Natl. Acad. Sci. USA.* 87:763-766.
- Navidi, W.C., G.A. Churchill, and A. von Haeseler. 1991. Methods for Inferring Phylogenies from Nucleic Acid Sequence Data Using Maximum Likelihood and Linear Invariants. *Mol. Biol. and Evol.* 8:128-143.
- Nei, M. 1987. *Molecular Evolutionary Genetics.* Columbia University Press, New York.
- Posten, M.O. 1989. An effective algorithm for the non-central chi-squared distribution function. *Amer. Statist.* 43:261-263.
- Rao, C.R. 1973. *Linear Statistical Inference and its Applications* 2nd ed. Wiley, New York.
- Shoemaker J.S. and W.M. Fitch. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. *Mol. Biol. and Evol.* 6:270-289.
- Woese, C.R. 1987. Bacterial Evolution. *Microbiology Reviews* 51:221-271.

TABLES

Table 1: Models fit to archaeobacteria data

Model	$G^2$	df	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$
Star	23.60	3	—	0.0630	0.0774	0.1823	0.2113	—
Tree 1	15.47	2	—	0.0615	0.0766	0.1621	0.1936	0.0335
Tree 2	23.10	2	—	0.0609	0.0742	0.1816	0.2099	0.0061
Tree 3	18.46	2	—	0.0532	0.0696	0.1786	0.2086	0.0196
Star	0.96	2	0.2927	0.0885	0.1134	0.2679	0.3098	—
Tree 1	0.90	1	0.3014	0.0898	0.1151	0.2773	0.3187	-0.0135
Tree 2	0.40	1	0.2997	0.0946	0.1237	0.2716	0.3156	0.0188
Tree 3	0.05	1	0.2845	0.0769	0.1049	0.2623	0.3049	0.0214

For each of the models that were fit to the rRNA data, we show the generalized likelihood ratio statistic  $G^2$ , its degrees of freedom and maximum likelihood parameter estimates  $\hat{\theta}_i$ . The unconstrained maximum value of the log-likelihood is  $-1908.563$ . All values were computed numerically using double precision arithmetic. Models without invariable sites are shown in the upper part of the table. Corresponding models with invariable sites are shown in the lower part of the table. Notice that for tree topology 1 with invariable sites, the estimated central branch length is negative.

Table 2: Sample sized for testing the Central Branch Hypothesis

$\theta_5$	<u>Topology 1</u>		<u>Topology 2 or 3</u>	
	$\beta:$ 0.60	0.80	0.60	0.80
0.10	0.80	1.30	0.41	0.87
0.05	3.43	6.90	1.40	2.37
0.02	22.42	54.58	7.78	12.47
0.01	91.04	236.94	29.76	48.01

Sample sizes are given in units of 1000 basepairs (kilobases). The outer branch lengths are  $\theta_1 = \theta_2 = 0.1$  and  $\theta_3 = \theta_4 = 0.3$ . For topology 1, the shorter branches are on the same side of the central branch and for topologies 2 and 3, the shorter branches are separated by the central branch. The proportion of invariable sites is assumed to be  $\theta_0 = 0.0$ . Sample sizes required for other values of  $\theta_0$  can be calculated as  $n/(1-\theta_0)$ .



### Figure Captions

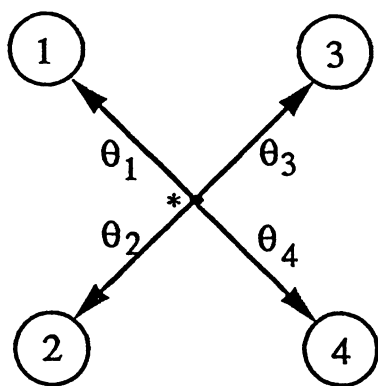
**Figure 1.** The star topology (a) and the tree topology 1 (b), represent alternative evolutionary relationships between four sequences. Two additional tree topologies can be obtained by relabeling the endpoints of the topology 1 tree.

**Figure 2.** Power for the test of the null hypothesis that the central branch has zero length ( $H_0 : \theta_5 = 0$ ) is shown for a tree with fixed and equal outer branch lengths  $\theta_1 = \dots = \theta_4 = 0.2$ , as a function of  $\theta_5$  under a model that does not include invariable sites (a) and as a function of  $\theta_0$  with the central branch length fixed at  $\theta_5 = 0.1$  (b). Individual curves are shown for different sample sizes  $n$  as indicated. Symbols are used to mark the power as estimated by simulations at  $n = 100$  ( $\square$ ),  $n = 500$  ( $\diamond$ ),  $n = 1000$  ( $\triangle$ ) and  $n = 5000$  ( $+$ ). The smooth curves in figures 2a and b (as well as in figures 3 and 4) represent asymptotic approximations which assume a large number of variable sites. Thus, the small sample curves tend to overestimate the actual power. This is also the reason for the convergence of power curves to  $1 - \beta = 0.05$  at  $\theta_0 = 1.0$  in figure 2b. The actual power at this point is zero.

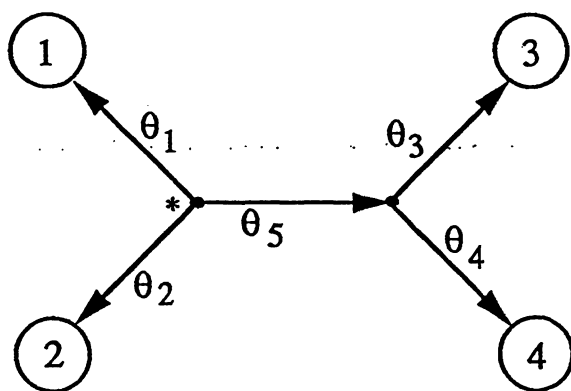
**Figure 3.** Power of the test of  $H_0 : \theta_5 = 0$  is shown as a function of sample size ( $n$ ) for trees with symmetric outer branch lengths equal to 0.1 (a), 0.2 (b) and 0.3 (c) and for trees with asymmetric outer branches arranged as follows:  $\theta_1 = \theta_3 = 0.1$ ,  $\theta_2 = \theta_4 = 0.3$  (d, short branches separated), and  $\theta_1 = \theta_2 = 0.1$ ,  $\theta_3 = \theta_4 = 0.3$  (e, short branches

together). The sample size axis uses a logarithmic scale. Individual curves are shown for central branch lengths  $\theta_5 = 0.1, 0.05, 0.02$  and  $0.01$  (left to right).

**Figure 4.** Power for the test of  $H_0 : \theta_5 = 0$  is shown as a function of  $\theta_5$  at sample size  $n = 1000$  for asymmetric trees with outer branches arranged as follows:  $\theta_1 = \theta_3 = 0.1, \theta_2 = \theta_4 = 0.3$  (a, short branches separated) and  $\theta_1 = \theta_2 = 0.1, \theta_3 = \theta_4 = 0.3$  (b, short branches together). Individual curves are shown for likelihood ratio tests based on binary data with no invariable sites (—), binary data with invariable sites  $\theta_0 = 0.0$  (....), quaternary data with no invariable sites (—.—.), quaternary data with invariable sites  $\theta_0 = 0.0$  (— — — —) and for the exact test based on Lake's invariant statistics (—...—). Notice that there is a slight loss of power when invariable sites are introduced into the models for binary and 4-state data even though the tree proportion is assumed to be zero. Notice also that the level of the invariants test is slightly less than  $\alpha = 0.05$ . This is because we have computed the exact power based on the binomial distribution (see Appendix B) rather than a large sample approximation.



**A**



**B**

